Building a Large-scale Persona Dialog Dataset

Yinhe Zheng[13], Guanyi Chen[2], Minlie Huang[3]
[1]Samsung R&D Institute of China - Beijing (SRC-B), Beijing, China
[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
[3]Conversational AI Group, AI Lab., Dept. of Computer Science, Tsinghua University
yh.zheng@samsung.com, g.chen@uu.nl, aihuang@mail.tsinghua.edu.cn

Abstract

We proposed a primary version of a large scale multi-turn dialogue dataset in Chinese that contains over 25 million sessions of dialogues crawled from Weibo[1]. Diversified *personality traits* for each dialogue participant are collected to facilitate modelling persona in dialogues. Our dataset fills the blank of the resources for building personalised dialogue systems in open-domain conversations and can also serves as an important resource for a wide range of studies.

1 Introduction

Creating natural or human-like interface is vital in recent advance of the research in humam computer interaction (HCI). Specifically, for natural language interaction, a human-like conversational agent is needed. Based on the theories in (computational) pragmatics or socialinguistics, people tend to perform specific personae when they produce language (Goffman, 1959; Shum et al., 2018). Therefore, one of the key feature of a human-like conversation agent is that it should be equipped with a personalised response generation system, i.e., it can generate coherent responses carrying different linguistic styles corresponding to diversified personality traits. Although there have a variety of neural models for dialogue generation. The studies regarding to personalised dialogue generation are still limited. The main reason is the lack of suitable large scale datasets that facilitate capturing general personae in dialogues.

This paper presents **PersonalDialog**: a large scale dialogue dataset collected from Weibo, which contains more than 25 million sessions of dialogues along with the rich structured personality traits of about 10 million speakers. These personal metadata not only contain the structured persona[2] information (which is similar to the key-value format used in Jurafsky et al. (1997) and Qian et al. (2017)), but also include the self-description of each speaker that are provided in natural language[3].

Note that in our daily life, dialogues are usually controlled by a mixture of three kinds of parameters: *content-based parameters* (e.g., aspect or dialogue act), *impersonal stylistic parameters* (e.g., politeness or tense), and *personalised stylistic parameters* (Ficler and Goldberg, 2017). This work focuses on building a large dataset and testing the controllability towards the personalised stylistic parameters. Actually, datasets used in previous works on modelling personalised dialogues are usually content-related, i.e., these works are restricted to a small domain and the persona are usually designed specifically for that domain (such as movies). Therefore, the resulting personalised dialogue generation model can only simulate certain extracted persona, which makes these models suffer from problems of sparsity and less controllability regarding to more generalised personae. In order to solve these issues, a large-scale dataset that contains rich structured personality traits is necessary to help modelling personae in open-domain dialogues, which is exactly what PersonalDialog provides.

In addition, previous studies also ignored an important phenomenon in modelling language production, that is, unlike content-based or impersonal stylistic parameters, which can always be

[1]Weibo (www.weibo.com) is one of the largest social media in China with hundreds of millions daily active users.

[2]Note that, in this paper, we define persona as a set of personality traits

[3]All the collected data are publicly available on Weibo and the information that can be used to back-trace the account of each user are not provided in PersonalDialog in order to protect the privacy of each user.

Workshop on Natural Language Generation for Human–Robot Interaction
at INLG 2018, November 8, 2018, Tilburg, The Netherlands.
http://purl.org/net/nlg-hri-workshop-2018/

expressed, people may not express their full-scale persona in every utterances they generate (Nguyen et al., 2014). Therefore, we argue here that a human-like personalised dialogue system should be able to **decide when and where to express which personality when generating responses with respect to the human input**. Previous datasets failed to handle such phenomenon since they don't have fine-grained personality traits provided. The structured personal metadata contained in PersonalDialog provides exactly what we need.

2 Dataset Construction

PersonalDialog is collected from Weibo with several elaborated strategies to facilitate the modelling of persona. Actually, people usually use Weibo to socialise and share feelings, and our initial observations suggest that the collected dialogues are usually of pretty high quality, especially for these dialogues that can last for several turns. Meanwhile, there are also various personality traits provided by users themselves on Weibo, which makes it an ideal source for building personalised dialogue datasets. Same to other user-generated corpora, our dataset also met the problem of noisy and sparsity towards the personality traits.

As for the issue of noisy, a two-stage data crawling schema is designed. Specifically, the first stage is about a careful seed users selection process, in which we manually select a number of News accounts who have a considerable number of followers, and collect users who comment under these news posted by those accounts . The second stage gathers all the dialogues under the weibo posted by these seed users together with their personality traits. About 50 million sessions of raw dialogues and 10 million of users' traits are collected. We believe our schema makes the crawler focusing on real active users rather than the water armies (Jindal and Liu, 2007) that are flooded on SNS.

After the crawling, a set of human defined filtering rules are designed based on various features: such as user levels and syntactic patterns. These features can be used to filter out the noisy users (e.g. spammers and bots) and posts.

As for the issue of sparsity, we decided to represent personality traits in our dataset as meta-data, in which each persona is stored as a set of key-value pairs, including the age, gender, user tags and personal descriptions extracted from user profiles. Comparing to the previous works that represent personae with contexts (Li et al., 2016; Zhang et al., 2018), our explicit representation can reduce the sparsity towards the personality traits.

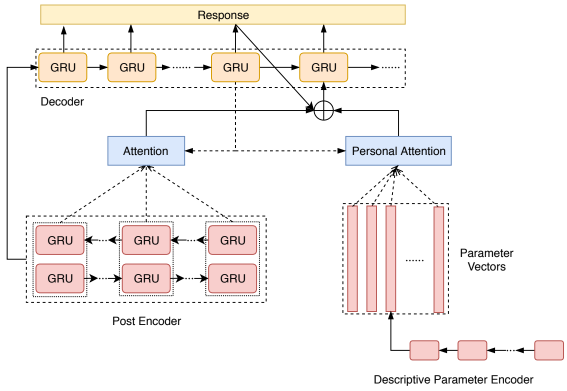

Figure 1: Sketch of our personalised dialog generation model.

3 Personalised Dialog Models and Dataset Evaluation

To test the usage of our corpus, we embedded persona attentions into the Seq2Seq framework (Bahdanau et al., 2014). The sketch of our model is shown in Figure 1. With a trait attention mechanism, we hope our model can some how learn to model persona based on contexts, i.e., the model can chose which persona to perform under different contexts.

Model	Perplexity	Gender Acc.	Age Acc.
Seq2Seq	84.07	50.2%	25.3%
Our Model	80.43	64.2%	42.6%

Table 1: Experiments results of dialogue generation models.

In order to evaluate whether the generated utterances indeed carry certain persona, we built two trait classifiers: a gender classifier and an age classifier. We tested a number of possible models for building classifiers. Considering the performance and efficiency, we finally chose FastText, which reaches an accuracy of 90.23% and 85.08% on the age and gender classification, respectively. We also evaluated different models using perplexity of the generated responses. Table 1 primarily compares the performance of the models with or without personal attention and Table 2 gives examples of responses generated by our model.

Test Post	Generated Response		English Translation
首先你要有一个男朋友 You should firstly have a boyfriend.	Male	我不要男朋友	I don't need a boyfriend.
	Female	你是我的男朋友	You are my boyfriend.
	Seq2Seq	我也这么觉得	I think so.
你怎么也这么多愁善感了? Why you are so sensitive?	Male	我是⟨UNK⟩	I am ⟨UNK⟩
	Female	因为我是小公举 公主	Because I am little princess.
	Seq2Seq	我一直都是	I am consistent.
把她电话号码给我 Please give me her telephone number.	Male	你要给我介绍吗	Will you introduce her to me?
	Female	你要给我打钱吗	It is not free!
	Seq2Seq	好的!	OK!
我一直都是你的小可爱 I am your little happiness forever!	Male	你是我的小可爱	You are my little happiness.
	Female	么么哒	XOXO!
	Seq2Seq	是呀是呀	Yes, Yes!

Table 2: Case study of the generated responses with respect to different setting of `gender`.

4 Conclusion and Future work

In the future, we tend to expend our work from the following aspects: 1) Apply more data denosing approaches, for example, using more detailed human defined filtering rules or directly modelling the noise in dialogue models using the techniques such as reinforcement learning; 2) Explore more objective evaluation metrics for personalised dialogue systems. One could design human evaluation schemes not only considering the quality of the generated responses but also considering whether the generated responses express the proper personal information. We plan to do a Turing test based on this scheme, that is, ask subjects to make judgements of whether the given responses generated by our models successfully simulate certain persona. Another evaluation scheme that should also be considered is the automatic evaluation metrics. A possible solution is to build more trait classifiers, and evaluate the responses using these classifiers. However, current experiments have shown that those classifiers are not friendly with models that are equipped with personal attention since the trait classifiers prefer models that can generate responses that always carry all the personalities, which is inconsistent with our initiation of building a human-like Chatbot. We also realised that our dataset can be applied beyond the task of building dialogue systems. We plan to use our dataset in the research areas like computational sociolinguistics (Nguyen et al., 2016) or social network analysis (Wasserman and Faust, 1994).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. *arXiv preprint arXiv:1707.02633*.

Erving Goffman. 1959. The presentation of self in everyday life. *New York*.

Nitin Jindal and Bing Liu. 2007. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 88–95. IEEE.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. pages 994–1003.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation.

Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?